rTASSEL: an R interface to TASSEL for association mapping of complex traits

Brandon Monier[1], Terry M. Casstevens[1], Edward S. Buckler[1,2]

1. Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853
2. United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853

Abstract

The need for efficient tools and applications for analyzing genomic diversity is essential for any genetics research program. One such tool, TASSEL (Trait Analysis by aSSociation, Evolution and Linkage), provides many core methods for genomic analyses. Despite its efficiency, TASSEL has limited means for reproducible research and interacting with other analytical tools. Here we present an R package rTASSEL, a front-end to connect to a variety of highly used TASSEL methods and analytical tools. The goal of this package is to create a *unified* scripting workflow that exploits the analytical prowess of TASSEL in conjunction with R's popular data handling and parsing capabilities without ever having the user to switch between these two environments.

Introduction

As breakthroughs in genotyping technologies allow for evermore available variant resources, methods and implementations to analyze complex traits are needed. One such resource is TASSEL (Trait Analysis by aSSociation, Evolution and Linkage). This software suite contains functionality for analyses in association studies, linkage disequilibrium (LD), kinship, and dimensionality reduction (e.g. PCA and MDS) (Bradbury *et al.*, 2007). While initially released in 2001, the fifth version, TASSEL 5, has been optimized for handling large data sets, and has added newer approaches to association analyses for many thousands of traits (Shabalin, 2012). Despite these improvements, interacting with TASSEL has been limited to either a graphical user interface with limited workflow reproducibility or a command line interface with a higher learning curve that can dissuade novice researchers (Zhang *et al.*, 2009). To remediate this issue, we have created an R package, **rTASSEL**. This package interfaces the analytical power of TASSEL with R’s data formats and intuitive function handling (R Core Team, 2019).

Approach

Implementation

rTASSEL combines TASSEL's abilities to store genotype data as half bytes, bitwise arithmetic for kinship analyses, genotype filtration, extensive forms of linear modeling, multithreading, and access to a range of native libraries while providing access to R's prominent scripting capabilities and commonly used Bioconductor classes (Gentleman *et al.*, 2004; Lawrence *et al.*, 2013; Morgan *et al.*, 2020). Since TASSEL is written in Java, a Java to R interface is implemented via the rJava package (Urbanek, 2019).

rTASSEL allows for the import, analysis, visualization, and export of various genomic data structures. Diverse formats of genotypic information can be used as inputs for rTASSEL. These include variant call format (.vcf), HapMap (.hmp.txt), and Flapjack (.flpjk.*). Phenotype data can also be supplied as multiple formats. These include TASSEL formatted data sets or R data frame objects (Figure 1A).

Once data is imported, an S4 object is constructed that is used for all downstream analyses (Figure 1B, 1C). To construct this object, the function, `readGenotypePhenotype`, is used. This S4 object contains slots that exclusively hold references to objects held in the Java virtual machine (JVM), which can be called with downstream functions. Prior to analysis, genotype objects can be filtered several ways to help in the reduction of confounding errors. rTASSEL can filter genotype objects by either variant site properties (`filterGenotypeTableSites`) or by individuals (e.g. taxa) (`filterGenotypeTableTaxa`).

Association functions

One of TASSEL's most powerful functionalities is its capability of performing a variety of different association modeling techniques. rTASSEL allows for several types of association studies to be conducted by using one basic function, `assocModelFitter`, with a variety of parameter inputs. This allows for implementing both least squares fixed effects general linear models (GLM) and mixed linear models (MLM) via the Q + K method (Yu *et al.*, 2006). If no genotypic data is provided to the GLM model, best linear unbiased estimates (BLUEs) can be calculated. Additionally, fast GLM approaches are implemented in rTASSEL which allow for the rapid analysis of many phenotypic traits (Shabalin, 2012).

The data model for an analysis can be specified by a formula similar to R's `lm` function (R Core Team, 2019) which is shown as follows:

$$y \sim A_1 + A_2 + ... + A_n$$

Where y is any TASSEL phenotype data and A is any TASSEL covariate or factor types. This formula parameter along with several other parameters allow the user to run BLUE, GLM, or

MLM modeling. Once association analysis is completed, TASSEL table reports of association statistics are generated as an R list which can then be exported as flat files or converted to data frames (Figure 1D). rTASSEL can also visualize association statistics with the function, `manhattanPlot`, which utilizes the graphical capabilities of the package, ggplot2 (Wickham, 2016) (Figure 1E).

Linkage disequilibrium

Linkage disequilibrium (LD) can also be generated from genotype data via the rTASSEL function, `linkageDiseq`. LD is estimated by the standardized disequilibrium coefficient, D', as well as correlation between alleles at two loci (r^2) and subsequent *P*-values via a two-sided Fisher's Exact test. TASSEL Table reports for all pairwise comparisons and heatmap visualizations for each given metric can be generated via TASSEL's legacy LD Java viewer or through ggplot2 (Figure 1F).

Relatedness functions

In order to perform MLM techniques, relatedness estimates (K) need to be calculated. TASSEL can efficiently compute this by processing blocks of sites at time using bitwise operations. rTASSEL can leverage this using the function `kinshipMatrix`, which will generate a kinship matrix from genotype data. Several methods for calculating kinship in TASSEL are implemented. By default, a "centered" identity by state (IBS) approach is used (Endelman and Jannink, 2012). Additionally, normalized IBS (Yang *et al.*, 2011) , dominance centered IBS (Muñoz *et al.*, 2014), and dominance normalized IBS (Zhu *et al.*, 2015) can be used. rTASSEL can either generate a reference object for association analysis or an R matrix object via the `kinshipToRMatrix` function for additional analyses.

Genomic prediction

In rTASSEL, the function `genomicPrediction` can be used for predicting phenotypes from genotypes. In order to do this, `genomicPrediction` uses genomic best linear unbiased predictors (gBLUPs). It proceeds by fitting a mixed model that uses kinship to capture covariance between taxa. The mixed model can calculate BLUPs for taxa that do not have phenotypes based on the phenotypes of lines with relationship information.

When the analysis is run, the user is presented with the choice to run k-fold cross-validation. If cross-validation is selected, then the number of folds and the number of iterations can be entered. For each iteration and each fold within an iteration, the correlation between the observed and predicted values will be reported. If cross-validation is not selected, then the original observations, predicted values and prediction error variance (PEV) will be reported for all taxa in the dataset.

Practical example

By using rTASSEL, end users can run various types of analytical pipelines with just a few functions. In the following example, single nucleotide polymorphism (SNP) data from the US nested association mapping panel in maize is used to analyze the trait “days to silk”.

For example, the following lines of R code will (I) import genotypic and phenotypic data, (II) calculate and create a kinship object, (III) calculate BLUEs and run MLM association analysis, (IV) generate a manhattan plot, (V) run genomic prediction, and (VI) generate an LD plot on a filtered genotype table using magrittr (Bache and Wickham, 2014):

```
## (I)
tasGenoPheno <- readGenotypePhenotype(
    genoPathOrObj    = “path/to/genotype.vcf”,
    phenoPathDFOrObj = “path/to/phenotype.txt”
)

## (II)
tasKin <- kinshipMatrix(
    tasObj = tasGenoPheno,
    method = “Centered_IBS”
)

## (III)
tasMLM <- assocModelFitter(
    tasObj     = tasGenoPheno,
    formula    = DaysToSilk ~ 1,
    fitMarkers = TRUE,
    Kinship    = tasKin
)

## (IV)
manhattanDTS <- manhattanPlot(
    assocStats = tasMLM,
    Trait      = “DaysToSilk”
)

## (V)
tasCV <- genomicPrediction(
    tasPhenoObj = tasGenoPheno,
    kinship     = tasKin,
    doCV        = TRUE,
    kFolds      = 5,
    nIter       = 1
)

## (VI)
# Require magrittr (%>%)
tasGenoPheno %>%
    filterGenotypeTableSites(
        siteRangeFilterType = “position”,
        startPos            = 228e6,
        endPos              = 300e6,
        startChr            = 2,
```

```
        endChr              = 2
    ) %>%
    ldPlot(
        ldType  = "All",
        plotVal = "r2",
        verbose = FALSE
    )
```

More information about various functionalities and workflows can be found on the online vignette (https://bitbucket.org/bucklerlab/rtassel/wiki/Home). Source code can be found on BitBucket (https://bitbucket.org/bucklerlab/rtassel/src/master/). An interactive Jupyter notebook session detailing additional rTASSEL workflows can be found on Binder (https://mybinder.org/v2/gh/maize-genetics/rTASSEL_sandbox/master?filepath=index.ipynb).

Acknowledgements

This project is supported by the USDA-ARS, the Bill and Melinda Gates Foundation, and NSF IOS #1822330. We thank Peter J. Bradbury and Guillaume Ramstein for their insightful suggestions on this manuscript and pipeline testing.

Figures

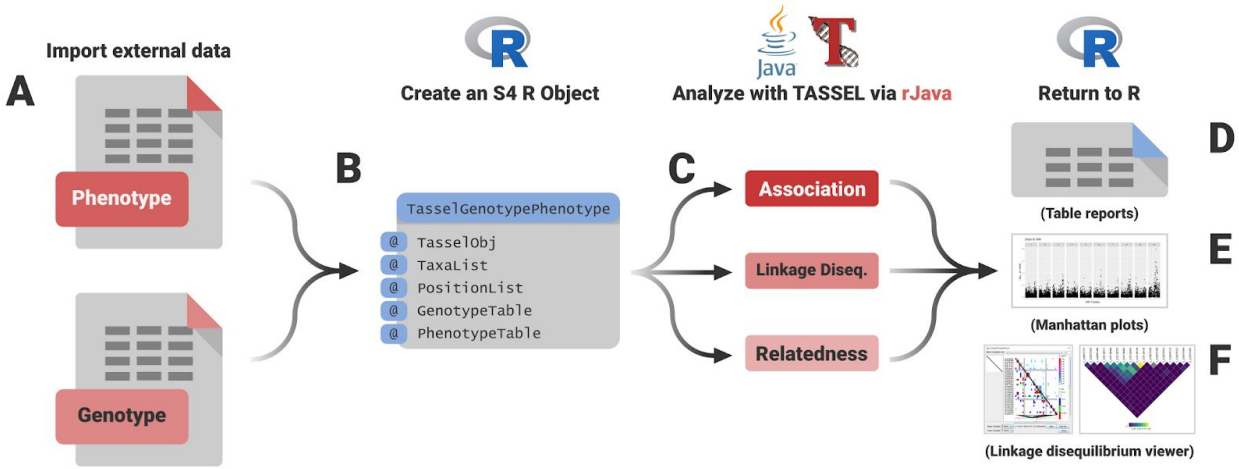


Figure 1: Overview of the rTASSEL workflow. Genotypic and phenotypic data (A) are used to create an R S4 object (B). From this object, TASSEL functionalities can be called to run various association, linkage disequilibrium, and relatedness functions (C). Outputs from these TASSEL analyses are returned to the R environment as data frame objects (D), Manhattan plot visualizations (E) or interactive visualizations for linkage disequilibrium analysis (F).

References

Bache,S.M. and Wickham,H. (2014) magrittr: a forward-pipe operator for R. *R Package Version*, **1**.

Bradbury,P.J. *et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Endelman,J.B. and Jannink,J.-L. (2012) Shrinkage Estimation of the Realized Relationship Matrix. *G3 Genes Genomes Genet.*, **2**, 1405–1413.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Lawrence,M. *et al.* (2013) Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.*, **9**, e1003118.

Morgan,M. *et al.* (2020) SummarizedExperiment: SummarizedExperiment container.

Muñoz,P.R. *et al.* (2014) Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. *Genetics*, **198**, 1759.

R Core Team (2019) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Urbanek,S. (2019) rJava: Low-Level R to Java Interface.

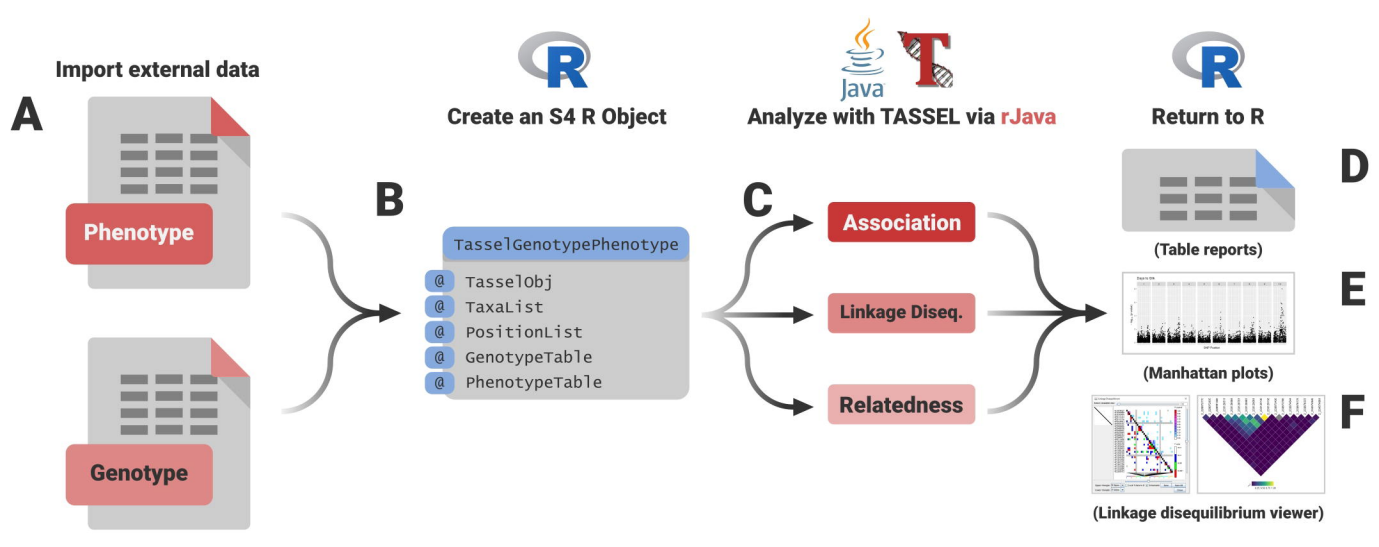
Wickham,H. (2016) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag New York.

Yang,J. *et al.* (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.*, **88**, 76–82.

Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Zhang,Z. *et al.* (2009) Software engineering the mixed model for genome-wide association studies on large samples. *Brief. Bioinform.*, **10**, 664–675.

Zhu,Z. *et al.* (2015) Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *Am. J. Hum. Genet.*, **96**, 377–385.


A
Import external data
Phenotype
Genotype
B
Create an S4 R Object
TasselGenotypePhenotype
@ TasselObj
@ TaxaList
@ PositionList
@ GenotypeTable
@ PhenotypeTable
C
Analyze with TASSEL via rJava
Java
Association
Linkage Diseq.
Relatedness
Return to R
D
(Table reports)
E
(Manhattan plots)
F
(Linkage disequilibrium viewer)